NCBI BlastRules Collection Release Notes

Version 4.0, March 8, 2021

BlastRules from NCBI is a set of manually created protein family classifiers, with attached metadata that automated annotation pipelines can use. Fields of interest always include the recommended protein product name, and may also include gene symbol, PubMed identifier (PMID), Enzyme Commission (EC) number, and free text comments about members of the BlastRule's family of matched proteins, or about the rule itself. The BlastRules database is designed to work together with other sources of evidence, such as collections of protein profile hidden Markov models (HMMs).

Automated prokaryotic genome annotation pipelines should be able to provide the best possible functional annotation for each protein by choosing the annotation rule of highest precedence from among all rules that cover that protein. BlastRules are a convenient way to build new annotation rules for families of proteins related to each other closely enough that BLAST's detection of pairwise sequence similarity, that basis of BLAST database searches, has sufficient sensitivity.

The required elements of each BlastRule are the rule's **accession** number, a **rule type**, a descriptive **protein name**, a list of one or more reference **proteins**, and a set of **threshold** values that help determine if a BlastRule hits a protein. Protein names are guaranteed to meet validation requirements for the submission of prokaryotic genomes to GenBank. Optional elements include a gene symbol, one or more PubMed identifiers, explanatory text, and EC numbers. Future releases will include Gene Ontology (GO) term identifiers. Rules may have additional information that is not provided in the current release, including tracking information such as owner and creation date.

BlastRules is used by NCBI's Prokaryotic Genome Annotation Pipeline (PGAP), its prokaryotic genome annotation tool. NCBI now provides a downloadable version of PGAP, at https://github.com/ncbi/pgap , with BlastRules included in the download package. The files included provided in this FTP directory are meant to support alternatives explorations or uses of BlastRules.

Provenance: The use of BlastRules in NCBI's prokaryotic genome annotation pipeline PGAP, or in the continually reannotated reference sequence collection RefSeq, is tracked. For new annotations by BlastRule, the source of the annotation is shown. See, for example, the Entry protein page for **WP_185168143.1**. The BlastRule **NBR012067** is cited on the protein page. It served as the source for the protein product name, the gene symbol, and citations to two publications about the protein.

WebPages for BlastRules. Each BlastRule is part of NCBI's collection of Protein Family Models, and soon will be searchable through NCBI's Entrez query system. The web giving details about a BlastRule, and providing lists of protein hit by that BlastRule, includes the BlastRule accession as

part of the URL. An example BlastRule web page is
https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/NBR012067/

Citing BlastRules: BlastRules was first described in Haft, et al., *RefSeq: an update on prokaryotic genome annotation and curation*, for the Nucleic Acids Res. database issue of January 2018. The article can be found at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753331/
After January 1, 2021, please check for availability of a newer publication on PGAP and RefSeq.

Release Schedule: The release schedule has not been determined. Release 3.0 reflects more than 16 months of additions and changes since release 2.0. NCBI expects future full releases to occur approximately once per year, as BlastRules enter production weekly within NCBI and do not depend the release cycle. NCBI will provide the means to run the collections of rule against a file of proteins in fasta format, and report which rules score qualifying hits to which proteins, at what precedence. Once the tool is available and in use, NCBI will increase the frequency of BlastRule database releases.

Release Statistics:

Release	Date	# of Rules	Chosen evidence	Total evidence
1.0	2017-09-15	840	63,644 proteins	162,890 proteins
1.1	2017-12-08	5568	120,786 proteins	222,497 proteins
2.0	2019-05-14	8159	684,140 proteins	728,208 proteins
3.0	2020-09-25	12,521	1,212,446 proteins	1,264,375 proteins
4.0	2021-02-25	13,381	1,331,330 proteins	1,389,770 proteins

Note that coverage of RefSeq proteins may be understated for some releases of NCBI BlastRules because of the time required to process, activate, and apply newly created rules.

Content and Coverage: BlastRules currently serve primarily as a mechanism to address limits in the expressive power of available annotation rules from other sources, such as CDD-SPARCLE architectures or *equivalog* HMMs. Most BlastRules so far, therefore, cover narrowly focused, high-interest, low-abundance proteins such as lineage-specific virulence proteins, serotype-specific markers, or vaccine candidate surface proteins from major human pathogens. Future waves of BlastRule creation, however, may address different sets of proteins and may change the overall character of database content and breadth of coverage by annotation pipelines. The largest single family of BlastRules is a set of over 4500 that name transposases by homology to those found in IS elements (insertion sequences – simple transposons that lack passenger genes) classified by ISFinder. These rules currently require 99% amino acid identity to annotate target proteins, and should help provide well-defined biomarkers useful for understanding chromosomal and plasmid evolution.

The content of the BlastRules database is provided by appropriately named columns in a tab-separated values file. The columns are as follows:

BlastRule_Acc: BlastRule accessions take the form NBRnnnnnn, where 'n' is any digit.

Rule type: BlastRules compete with each other, and with other types of evidence, for the right to determine the name a protein should receive. The rule type determines precedence, on an arbitrary scale. So far, three types of rules are defined. The rule type also determines default values for three threshold parameters. *Identity* is the threshold for the percent identity between a BlastRule reference protein and candidate matching sequence, ignoring gap regions. *Model coverage* is the percent of the length of the reference sequence that must be aligned for a hit to be recognized. *Target coverage* is the percent length of the sequence being tested for a match to the rule. The table below shows the **default** cutoffs of amino acid percentage identity, and coverage sequence length in alignment by BLAST, for six types of BlastRules in descending order by precedence.

Rule Type	**Amino acid % identity**	**Model protein % coverage**	**Target protein % coverage**	**Typical use**
BlastRuleIS	99	95	90	transposase
BlastRuleException	94	90	90	virulence factor
BlastRuleEquivalog	80	90	80	enzyme
BlastRuleSubPlus	60	85	85	various
BlastRuleSubMinus	35	80	80	various
BlastRuleCOLLAB	95	90	90	importing rules

The curator of a given BlastRule can change these thresholds from the initial (default) cutoffs as needed for that rule.

Note that the term "*equivalog*" describes a set of proteins that share a specific function by virtue of evolutionary descent from an ancestral sequence with that same function. *BlastRuleException* rules attach names that are even more specific than *equivalog* names, as when the literature distinguishes among different isozymes from the same *equivalog* family, or among closely related virulence proteins.

Name: Names from BlastRules are designed to comply with the standards that GenBank and RefSeq require for valid protein names, and to become the protein product name in PGAP and RefSeq annotations. A typical name, "tandem repeat protein effector TRP47", follows the literature as closely as it can, but avoids a discouraged and troublesome explicit reference to molecular weight, "47 kDa." Proteins recognized by the rule, in fact, are repeat-rich and quite variable in length, and the name "TRP47" remains appropriate irrespective of the computed molecular mass.

GB_proteins: This column (often empty) contains a column-separated list of accession numbers for proteins cited as a reference sequences for a BlastRule for all cases in which the accession number is taken from GenBank, SwissProt/UniProt, of a RefSeq NP or YP series protein. The next column contains WP-series accessions for all BlastRule proteins. A single BlastRule may be defined by several proteins in order to minimize the risks of having to use one protein only but overly permissive cutoffs.

WP_proteins: This column provides WP-series accession numbers all non-redundant proteins used in BlastRules, as a comma-separated list. It includes proteins 100% identical to any non-WP proteins cited in the previous column.

Identity: A BlastRule hits a target protein if three conditions are met: percent identity, percent of the BlastRule's model protein aligned to the model by BLAST, and percent of the target protein's length aligned to the model. This column contains the percent identity cutoff, typically 94% for *BlastRuleException*, 80% for *BlastRuleEquivalog.* In release 3.0, the lowest percent amino acid sequence identity required by any BlastRule is 45 percent.

Model_pct: The protein(s), used to define a BlastRule is (are) the model(s) for that rule. Model_pct is simply the miminum percent of a model protein's length that must be used by BLAST to align to a target protein that the BlastRule identifies as a match. Post-processing of BLAST alignments is presumed, so two or more hit regions can be combined as long as no region of either the model protein or the query protein can be used twice and aligned regions are in the same order on both proteins.

Target_pct: This number is the miminum percent of a target protein's length that must match to a BlastRule model protein for evidence of a BlastRule hit to register. It measures on a target protein what Model_pct measures on a BlastRule model protein. Note that we refer to Model and Target, rather than Query and Subject, because the choice of which sequences to treat as the query and which to make subjects in a BLAST-searchable database may not always be obvious.

Gene: This optional field must have a single value only that complies with standards for GenBank's standards for gene symbols in prokaryotic genomes, or else is left null.

PMID: PubMed identifiers describe a BlastRule on the whole, although users may be able to infer how individual PMID link to individual model proteins.

EC: Enzyme Commission numbers. May be a comma-separated list.

Comment: Each rule may have a plain text comment consisting of explanatory text that eventually could accompany the public annotation of a protein, as through the Entrez query system. The comment must be formatted as a single line (no carriage returns or line feeds allowed).

For Researchers with Highly Curated Specialty Sets: Currently, BlastRules are used only for prokaryotic annotation. NCBI is interested in obtaining new sources of BlastRules from experts in specialized collections of prokaryotic proteins, such as bacteriocins or immunity proteins discussed in review articles by those experts. If interested in seeing BlastRules created, and RefSeq annotation properly updated, for a collection of well-understood proteins, please contact the NCBI HelpDesk or the corresponding author of PMID:29112715.